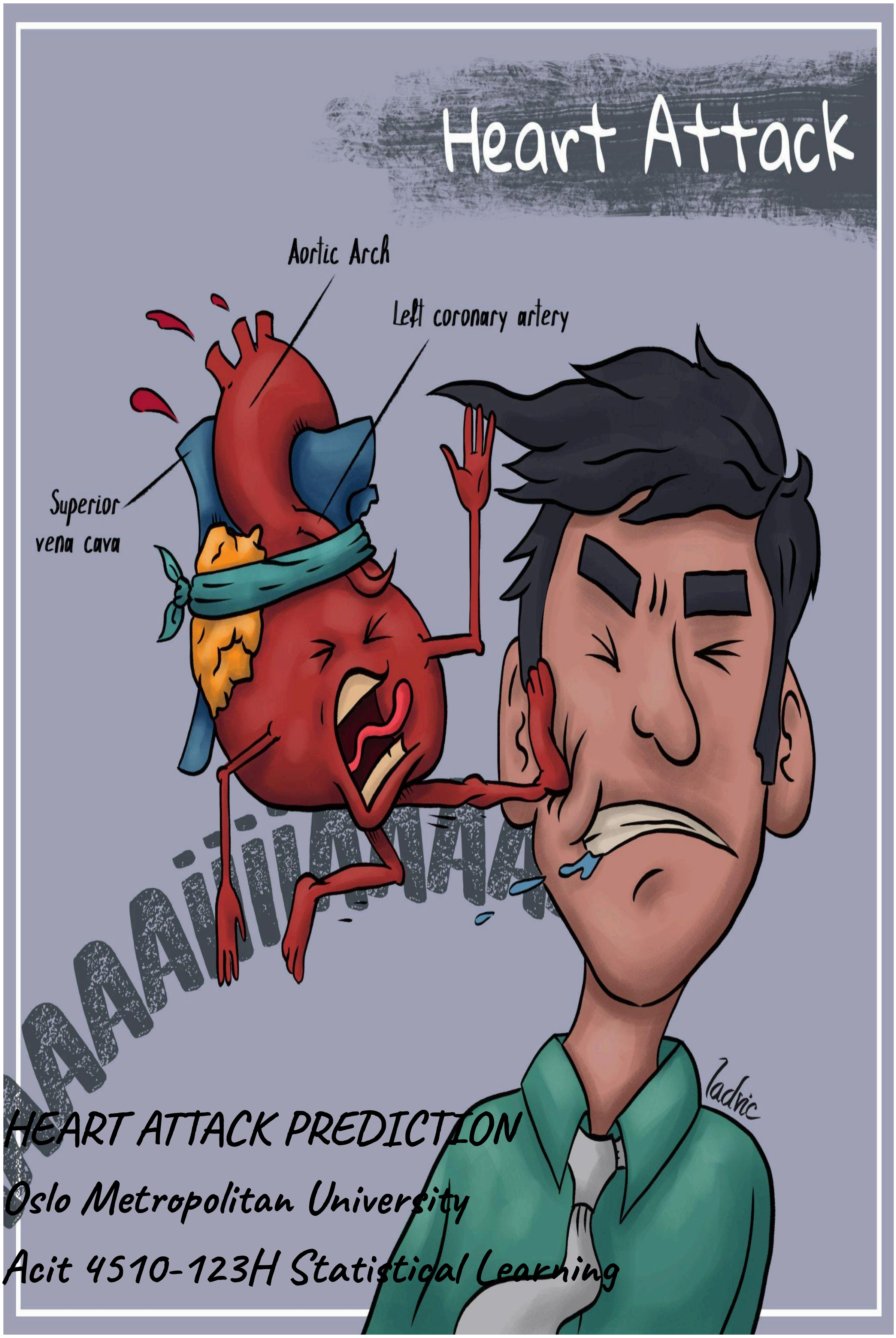


HEART ATTACK PREDICTION

Oslo Metropolitan University

Acit 4510-123H Statistical Learning

Table of contents

1.Abstract

Over the past few decades, Heart attack problem has become one of the most dangerous threats for human beings , being a leading cause of death globally. Due to that alarming situation, Researchers community proposed some machine learning techniques in order to address the pressing need for timely and accurate diagnosis in order to facilitate appropriate treatment. The technology improvement allows the ability to automatically analyze large and complex medical data sets. Consequently, many studies have been carried out to help medical professionals diagnose heart-related diseases. In this paper, a variety of machine learning algorithms are surveyed and their performance is evaluated. Those models are supervised learning models such as Random Forest (RF), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression and unsupervised models such as PCA.

2.Introduction

Myocardial infarctions which are known as heart attacks are a major global health concern that have significant effects on not only individuals but also healthcare systems. In a time when heart attacks are still the primary cause of death, it is more important than ever to understand and forewarn heart attacks. When blood flow to a portion of the heart muscle is cut off, a heart attack happens, frequently causing irreversible damage. Beyond the acute health effects, such an incident has a negative impact on people's quality of life and strains worldwide healthcare systems. Being aware of how urgent it is to solve this problem, scientists and medical professionals are always looking for new and creative ways to anticipate and prevent heart attacks.

Predicting heart attacks early on is essential to reducing their catastrophic consequences. Proactive measures that have the potential to prevent or minimize the severity of an incident are made possible by the timely identification of those who are at risk. Reliability prediction models are crucial for personalized treatment and targeted interventions as the medical community places a greater emphasis on preventative healthcare.

Heart attack prediction techniques now in use include a variety of risk assessment models and diagnostic instruments. Current approaches provide excellent insights into a variety of risk variables, from age and family history to contemporary imaging techniques and biomarker analysis. Nevertheless, issues still exist, such as poor precision and the incapacity to fully represent the intricacy of the multivariate risks connected to heart attacks.
By examining new risk factors which are more relevant to medical measurements than demographics factors and taking advantage of predictive analytics , this paper aims to close the current gap in heart attack prediction approaches. By doing this, we want to improve the accuracy and usefulness of predictive models and advance our knowledge of the dynamics underlying the occurrence of heart attacks.

The question has been raised about how important this research is to the community? This work is important because it has the potential to completely change the way that cardiovascular health is currently practiced. Better heart attack prediction is a key component of larger public health programmes aiming at lowering the worldwide burden of cardiovascular diseases, as well as a promising avenue for tailored patient therapy.

The synopsis of the report is structured as follows: It starts with a thorough review of the literature, then moves on to a section on methodology, presents and analyzes the results, and ends with recommendations for clinical practice and future paths in research.All of the main points will be examined in more details in the section that follow ,giving a thorough analysis of heart attack prediction and its public health consequences.

3.Data description

3.1.Overview about the data and variables

The dataset obtained from Kaggle for predicting heart attacks consists of 1025 instances and 14 attributes. The primary objective is to utilize machine learning models to anticipate heart attacks based on various health-related features. The dataset includes crucial indicators such as age, which provides a continuous representation of individuals' years, and gender, denoted by a binary classification. The classification of chest pain type introduces a categorical dimension, categorizing it into typical angina, atypical angina, non-anginal pain, and asymptomatic.

Important health metrics such as resting blood pressure, serum cholesterol levels, and maximum heart rate achieved are captured as continuous variables. Fasting blood sugar levels are binary, aiding in the identification of potential diabetic conditions. The results of resting electrocardiograms provide categorical insights into normalcy or abnormalities. Exercise-induced angina, another binary attribute, indicates the presence or absence of angina during exercise.

The dataset further explores the complexities of health with features like ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, and the number of major vessels colored by fluoroscopy, providing additional layers of information. The target variable, which is denoted as “target” in the heart.csv file, encapsulates the presence or absence of disease, especially in this case is heart attack.

3.2.Quality Assurance of the Data

It is worth noting that the dataset does not contain any missing values, which attests to its reliability.Even though, the dataset's remarkable lack of missing values is an important aspect to consider, it is equally crucial to acknowledge the presence of duplicated values, as they can significantly impact the overall reliability of the information. Upon meticulous examination, it was discovered that out of the 1025 instances, a substantial 723 records are duplicated, which represents a noteworthy portion of the dataset. While this duplication does not necessarily invalidate the dataset, it highlights the significance of performing data cleaning and preprocessing steps to ensure the accuracy of predictive models. In addition to the presence of duplicates, it is worth noting that the dataset remains viable for machine learning applications.

Despite the dataset's moderate size after removing duplicates , which consists of 302 observations and 13 variables, it remains viable for conducting an analysis aimed at predicting heart attacks using machine learning methodologies. Although a larger dataset is generally preferred, the current size is not prohibitive, particularly for algorithms like logistic regression that can perform adequately with smaller datasets. The 13 variables encompass a range of health-related features, including age, gender, chest pain type, and cholesterol levels, providing a sufficiently rich set of attributes for constructing predictive models. The target variable's binary nature, indicating the presence or absence of heart disease, aligns well with the requirements of various machine learning algorithms. While the dataset size may present some limitations, especially for more complex algorithms like support vector machines, careful consideration of model complexity and thorough data preprocessing can help mitigate potential challenges. Implementing techniques such as cross-validation allows for a robust assessment of model performance, offering insights into the reliability of predictions. Additionally, given the constraints, opting for simpler models with the size of dataset (302,13) that are less prone to overfitting is a pragmatic approach. In summary, while acknowledging the dataset's moderate size, it is considered suitable for analysis, provided that appropriate considerations are taken into account during the modeling process.

3.3.Ethical consideration

The utilization of health-related datasets, particularly those containing sensitive information like heart health, raises significant ethical concerns such as medical identity theft or privacy violation. As we delve into health attributes such as chest pain type, blood pressure, and cholesterol levels, it is important to recognize the potential sensitivity of this information. The significance of this dataset in predicting heart attacks lies in its comprehensive coverage of various health indicators. The inclusion of age, gender, chest pain type, and other critical variables aligns perfectly with the objectives of our analysis. The dataset's wide range of features provides a holistic understanding of an individual's health, enabling a detailed exploration of factors contributing to heart attacks.

Adhering to data protection regulations and ethical guidelines is essential to prevent any potential misuse of personal health information
It is crucial to handle such datasets responsibly and ethically. In the case of this heart attack prediction dataset sourced from Kaggle named “heart-cvs”, a nuanced approach is necessary to ensure privacy and confidentiality. Responsible data usage involves implementing strict security measures, including encryption and restricted access, to prevent unauthorized disclosure.The utmost priority is to ensure the privacy and anonymity of individuals represented in the dataset . In the case of this analysis, the

“heart-csv” dataset did not apply encryption but at least all the observations have removed identity information and personal identity is not traceable, partly aligned to the moral assurance when working with health data.

4.Statistical inference

4.1.Descriptive Statistics

The dataset's composition is revealed through descriptive statistics, providing valuable insights. The average age of individuals in the sample is 54.42 years, indicating a mature population with the potential to have heart attacks . With a relatively low standard deviation of 9.05 years, there is a moderate spread, suggesting a focus on a specific age range. The dominance of males, constituting 68.21%, highlights a gender imbalance that should be considered in further analyses to avoid bias parameters. The prevalence of chest pain types 0 and 1 implies a significant proportion of individuals experiencing typical and atypical angina. Resting blood pressure and serum cholesterol levels serve as essential cardiovascular health indicators, with means of 131.60 mmHg and 246.50 mg/dL, respectively. Additionally, the presence of fasting blood sugar levels above 120 mg/dL in 14.90% of cases signals a potential diabetic predisposition in the dataset. Most individuals exhibit normal resting electrocardiogram results (category 0), indicating a baseline cardiac health status. The mean maximum heart rate achieved during exercise is 149.57 beats per minute, with a notable range from 71 to 202 beats per minute, highlighting the variability in cardiovascular responses. Approximately 32.78% of individuals experience exercise-induced angina, reflecting a significant subset with heightened cardiovascular reactivity. The mean ST depression of 1.04 mm, ranging from 0 to 6.2 mm, suggests varying degrees of ischemia.

4.2.Correlation analysis

In this examination of heart attack prediction, correlation is employed as a statistical deduction tool with various objectives. It serves to identify multicollinearity when two independent variables are highly correlated and offers valuable insights into the tendency of paired variables' relationships.

First of all, to identify potential multicollinearity, a correlation matrix can be examined to visually determine if any variables are strongly correlated with each other.

In the heatmap of the correlation matrix, we can see that “oldpeak” and “slope” indicate some level of multicollinearity even though the correlation is not extremely high.

Secondly, the analysis of the correlation matrix uncovers interesting connections between important attributes in the dataset. It is noticeable that age has a negative correlation (-0.395) with maximum heart rate during exercise (thalach), which suggests that older individuals may have a lower maximum heart rate on average. Additionally, the positive correlation (0.432) between chest pain type (cp) and the target variable implies a possible link between the type of chest pain and the likelihood of heart disease. Furthermore, the positive correlation (0.287) between exercise-induced angina (exang) and ST depression induced by exercise relative to rest (oldpeak) suggests that the presence of exercise-induced angina is associated with a higher degree of ST depression. Lastly, the strong negative correlation (-0.576) between the slope of the peak exercise ST segment and oldpeak indicates that a downsloping slope is linked to a higher ST depression .

In terms of extreme cases, one of the strongest positive correlations (0.384754) is observed between the maximum heart rate achieved (thalach) and the slope of the peak exercise ST segment (slope). This indicates that individuals who achieve a higher maximum heart rate during exercise are more likely to have a steeper slope on their electrocardiogram during peak exercise. On the other hand, a significant negative correlation (-0.377411) is discovered between the maximum heart rate achieved (thalach) and the presence of exercise-induced angina (exang). This suggests that individuals who achieve a higher maximum heart rate during exercise are less likely to experience angina, indicating a potential protective effect of higher heart rates during physical activity. Furthermore, a weak positive correlation (0.011428) is observed between fasting blood sugar levels (fbs) and serum cholesterol levels (chol). Although this correlation is close to zero, it implies a subtle positive relationship between these variables. However, it's important to note that the correlation is weak, indicating that fasting blood sugar and serum cholesterol levels may largely operate independently of each other in this dataset.

It would be useful to narrow down the direct correlation of 13 variables to target one. In table 2, the strongest positive impact (0.432080) is observed in cp (chest pain type), indicating that a higher value is linked to a higher likelihood of heart disease. Conversely, exang (exercise-induced angina) demonstrates the strongest negative impact (-0.435601), implying that a higher value is associated with a lower likelihood of heart disease. These results emphasize the significant role of chest pain type and exercise-induced angina in predicting the presence of heart disease. These discoveries offer valuable insights for selecting features and further exploring the interrelationships within the dataset.

Correlation heatmap has been attached as Image 1 in Appendix

4.3. Variance Inflation Factor

Due to the unclear signals for multicollinearity from moderate correlation values ,VIF will be considered as a formal test.

The Variance Inflation Factor (VIF) is a metric utilized to evaluate the degree of multicollinearity . It measures the extent to which the variance of an estimated regression coefficient increases when the predictors are correlated. High VIF values suggest the possibility of multicollinearity, which can have an impact on the reliability (p-value) and interpretability of regression outcomes.

VIF could be calculated by :

$$VIF_i = \frac{1}{1-R_i^2}$$

where R square is measurement of proportions of the variance in the dependent variable that is explained by the independent variables In this analysis, results of the VIF test are shown in table 3.

This table indicates the presence of multicollinearity among the predictor variables . Variables with VIF values greater than 10, such as age, trestbps, chol, thalach, slope, and thal, suggest potentially severe multicollinearity, indicating a high correlation between these variables and others in the dataframe. The variable slope falls into a moderate range, which suggests a moderate level of correlation with other predictors . Variables with VIF values below 5, such as sex, cp, fbs, restecg, exang, oldpeak, and ca, indicate lower levels of multicollinearity. Those high VIF values suggest that it potentially impacts the reliability of their coefficient estimates. Variables with moderate or low VIF are generally considered acceptable in terms of multicollinearity.In order to tackle the issue of multicollinearity for variables with high VIF, it is important to explore various strategies such as implementing regularization techniques like PCA. This approach will be incorporated into the analysis.

VIF results has been shown in table 1 of Appendix

4.4. Examination of outlier detection and distribution for continuous variables.

The variable "Age" exhibits a bell-shaped distribution, indicating a relatively equal distribution of individuals across different age groups. On the other hand, "Trestbps" (Resting Blood Pressure) displays a distribution that is skewed to the right, suggesting that most individuals have lower resting blood pressure values, while fewer individuals have higher values. Similarly, "Chol" (Serum Cholesterol) also shows a right-skewed distribution with a long right tail, indicating that the majority of individuals have lower cholesterol levels, but there are only some individuals with extremely high values from 500 to 600. In contrast, "Thalach" (Maximum Heart Rate Achieved) demonstrates a distribution that is skewed to the right, implying that most individuals achieve higher maximum heart rates, with fewer individuals reaching lower values. Furthermore, "Oldpeak" (ST Depression Induced by Exercise Relative to Rest) presents a distribution with a left tail and a long right tail, indicating that most individuals experience lower levels of exercise-induced ST depression, but some individuals have higher values. The presence of a significant number of values at 0 in "Oldpeak" suggests that a notable proportion of individuals do not experience ST depression relative to rest. It is important to note that the skewness observed in these variables can have an impact on statistical measures such as the mean, median, and standard deviation, which in turn can influence the interpretation of these measures in this data analysis project. To address both the skewness and outliers observed , appropriate transformations such as Robust Scaler packages from Scikit-learn may be considered to improve the accuracy.

Histograms of numericals variables has been attached as Images 2 of Appendix

4.5. Verification of data imbalance.

4.5.1 Independent variables

The distribution of categorical variables can be effectively visualized through bar charts, offering valuable insights into key characteristics about data imbalance within the dataset.Evidently, the gender distribution reveals that approximately 68.2% of individuals are men, indicating a moderate imbalance in gender representation. When examining the types of chest pain (CP), it becomes apparent that 47.4% of individuals experience typical angina. On the other hand, there is a significant imbalance in the levels of Fasting Blood Sugar

(FBS), with 85.1% of individuals unlikely to be diabetic. Additionally, the distribution of Exercise Induced Angina (Exang) indicates that around 67% of individuals do not experience this condition, suggesting a notable imbalance in favor of those without exercise-induced angina. Within the dataset, only 1.3% of individuals exhibit probable or definite left ventricular hypertrophy by Estes' criteria, making this condition a minority and potentially posing challenges for model training and generalization due to its limited representation. Similarly, a small proportion (1.3%) of individuals have coronary arteries supplying blood to the heart tissue, further highlighting an imbalance in the distribution of this specific cardiac vessel condition. Imbalances in the independent variables, although they may have implications for comprehending the dataset or specific characteristics, might not have a direct influence on the predictive accuracy of the models used in this analysis, particularly if the target variable is quite evenly distributed.

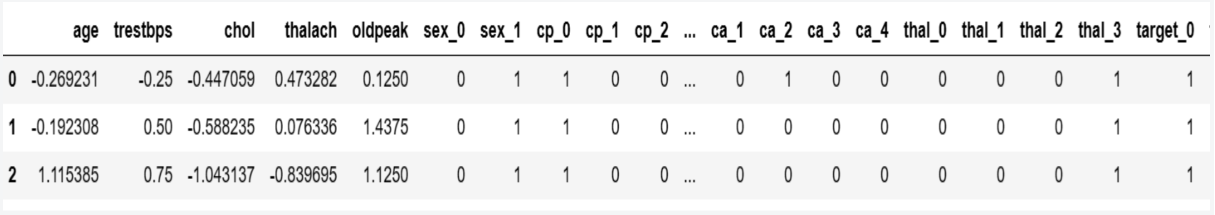
4.5.2 Dependent variable

The targeted variable exhibits a slight imbalance, with 54.3% of samples indicating heart disease and 45.7% showing no disease. Because this imbalance is considered negligible, the resampling technique will not be used. One approach to tackle this issue is by exploring the use of class weights during model development, particularly in decision trees technique. When dealing with imbalanced classes, it is crucial to consider significant metrics like precision, recall, and F1 score for model evaluation, rather than relying solely on accuracy metrics. These metrics provide valuable insights into the performance of the model for minority classes. In the upcoming section dedicated to comparing model performance, these metrics will be thoroughly examined.

5. Data processing

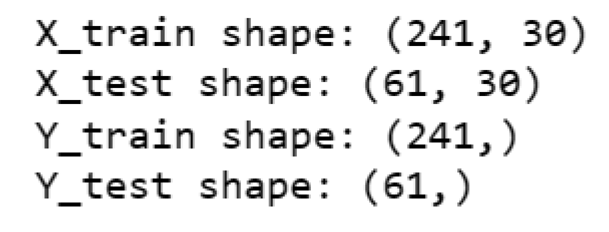
During the early stages of this heart attack prediction project, meticulous attention was paid to the preprocessing of categorical variables, a critical step to improve the model's interpretability and learning ability.

Using the pd.get_dummies method, categorical data are transformed into dummy variables. This technique efficiently encodes qualitative information in binary format, providing a structured representation for easy model understanding. For example, the resulting processed dataset, visualized using head(), revealed transformation, with columns like "sex_0", "sex_1", "cp_0", "cp_1", etc and so forth, capturing the encoded categorical information.



	age	trestbps	chol	thalach	oldpeak	sex_0	sex_1	cp_0	cp_1	cp_2	...	ca_1	ca_2	ca_3	ca_4	thal_0	thal_1	thal_2	thal_3	target_0
0	-0.269231	-0.25	-0.447059	0.473282	0.1250	0	1	1	0	0	...	0	1	0	0	0	0	0	1	1
1	-0.192308	0.50	-0.588235	0.076336	1.4375	0	1	1	0	0	...	0	0	0	0	0	0	0	1	1
2	1.115385	0.75	-1.043137	-0.839695	1.1250	0	1	1	0	0	...	0	0	0	0	0	0	0	1	1

After this categorical transformation, a critical standardization procedure was applied to numerical features. It was decided to use the scikit-learn RobustScaler in light of the identification of skewed data and possible outliers of the variables such as ‘thalach’, ‘oldpeak’. The RobustScaler makes use of percentile-based statistics, particularly the median and interquartile range (IQR), in contrast to the StandardScaler, which presupposes a normal distribution and is susceptible to outliers. This improves the scaling process' resistance to outliers, enabling the model to handle skewed data distributions more effectively.



```
X_train shape: (241, 30)
X_test shape: (61, 30)
Y_train shape: (241,)
Y_test shape: (61,)
```

The final dataset was now ready for further analysis after being enhanced with dummy variables and standardized numerical features using RobustScaler. In addition to guaranteeing that the model is compatible with machine learning algorithms, this careful data processing lays the groundwork for reliable model training and evaluation.

In the next stage, the dataset is strategically split into training and test sets using scikit-learn's train_test_split function.
Using a stratification method to maintain the distribution of the target variable, 80% of the data is allocated for training, while 20% is reserved for testing.
The shapes obtained from the training and test sets are as follows:

This meticulous data processing not only ensures model compatibility with machine learning algorithms but also paves the way for robust model training and evaluation, thereby creating a solid basis for accurately predicting heart attacks.

6.Model Selection and Analysis Strategy

6.1.Methodology Description

6.1.1. Decision Tree.

Decision trees are commonly used in education,healthcare,finance and other areas. Basically, decision tree is a data mining tool which is used to classify and forecast data.In contrast of black-box model where internal working is not visible and unstandable, a white-box system is implemented for Decision Tree model .It brings the advantage of being transparent and interpretable mechanisms that is rational behind model's predictions. This is one of certain types of rule-based systems where users can observe the direct trace how input features contribute to the final model's output

6.1.1.1. Introduction decision tree principles and kinds

Decision tree is graphically represented as a flow-chart tree structure where node represents attest on attribute value, branches represents test outcomes and leaves represents class distribution.There are two types of trees.In the context of our analysis for heart disease prediction , classification tree could be applicable.

6.1.1.2.Describe basic algorithms

Decision Tree algorithms build decision trees from datasets in one of two ways: top-down or bottom-up. Growth and pruning phases are performed by top-down algorithms, but some algorithms simply execute the growing phase. After the tree is built from the top down, it is pruned from the bottom up to solve overfitting.

6.1.1.3. Decision tree creation with fundamental points and factors that affect tree performance.

Firstly, Decision trees use entropy , information gain ratio and Gini Index to choose optimal property as splitting criteria for the tree.In this analysis, Gini index is mainly used for measuring the impurity of a dataset, representing the divergence between probability distributions of target attribute values.This factor plays crucial roles in determining how decision trees are constructed and how they handle different types of data in the context of medical dataset.

Secondly, the algorithm take into account the data types attribute for choosing properly generated model.For example, If it detects continuous variables, regression tree should be proposed to be chosen instead of classification trees .Moreover , the overfitting issues will be tackled by utilizing pruning strategies to reduce data misclassification.

6.1.2. Random forest

Random forests, which were introduced by Breimen in 2001, stand out as powerful ensemble learning algorithms designed to improve classification accuracy. It uses multiple classification trees in groups, which can be called decision tree bootstrapping. In detail, this ensemble of tree-based classifiers is created by generating trees from bootstrap samples and replacing them with calibration data. This algorithm uses a maximum voting rule to assign pixels, which is final classification for a particular instance, to a particular class based on the majority of votes received from an ensemble of classification trees.

A distinguishing feature of random forests is their robustness to overfitting. This generates a large number of trees and ensures generalization of the pattern to unseen samples. This algorithm requires adjusting only two parameters: the number of trees and the number of features/split variables.

Random forests estimate classification accuracy by using out-of-bag samples that are not part of the training set for each tree.

6.1.3 Bagging

Bagging is another ensemble learning technique introduced by Breiman in 1996. This method shares a common goal of improving classification accuracy.
Similar to random forests, bagging involves a set of tree-based classifiers generated by a bootstrap technique that randomly samples and replaces a portion of the calibration data. Nevertheless, unlike random forests, bagging considers all feature variables as candidates for splitting at each node.
This makes bagging easier in terms of parameter tuning, we only need to specify the number of trees.
Bagging indicates the stability and robustness of the calibration data to noise. Similar to random forests, bagging relies on an ensemble of trees and is therefore resistant to overfitting.The choice between two methods is remarkably dependent on consideration of researcher's preferences for dataset characteristics and ease of implementation.

6.1.4 Boosting

Boosting is the other advanced method rooted from Decision Tree which was developed by Freund and Schapire in 1996. It takes a different approach of converting a group of weak classifiers into strong classifiers, ultimately improving classification accuracy. Unlike random forests and bagging, boosting uses the entire calibration data set for classification without resampling.
It iteratively adjusts the weights based on their fit to the calibration data, and samples that were incorrectly classified in previous iterations are given higher weights.
Boosting prioritizes samples that were misclassified in previous iterations, effectively rectifying classification errors by assigning greater importance to those particular samples.

This iterative process results in an improved classifier that takes advantage of the strengths of each iteration. In particular, boosting requires adjusting the number of boosting iterations as the main parameter.
In conclusion, when random forests gain benefits of robustness and resistance, bagging demonstrates stability, boosting improved classification accuracy that is iteratively refined by focusing on samples that were misclassified in previous iterations.

In this analysis for heart attack prediction, all 3 methods will be applied along with parameter tuning complexity adjustment to see how good the testing accuracy improvement is.

6.1.5 . K-nearest neighbors algorithm

The K-Nearest Neighbors (KNN) algorithm, classified as lazy learning, selects neighbors from a data set with similar attributes or values.
The algorithm uses Euclidean distance, which represents a straight path between two points. The Euclidean distance between points A and B is calculated using the formula:

$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_2 - y_1)^2}$$

In this formula , d(A;B) denoted as the Euclidean distance between points A and B with pointA (x1,y1) , point B(x2,y2)
KNN algorithm starts with selecting the data set, randomizing, normalizing to scale the parameter values, creating a training model, and finally evaluating the accuracy of the algorithm's predictions.
Before applying KNN, preparing the dataset is very important, involving scaling parameters within the normalization range. Theoretically, KNN relies on distance calculations and normalization is to prevent larger scale features from dominating the results. It promotes equal feature importance, faster

convergence, consistent and robust comparisons, thereby improving model accuracy and generality to new data.

In this analysis, normalization techniques: Standard Score (Z-Score) normalization will be used . Z-Score normalizes each score based on its standard deviation from the mean. The Z-Score normalization formula is given by the formula :

$$\frac{X-\mu}{\sigma}=\frac{X-mean}{standarddeviation}$$

6.1.6 Principal component analysis

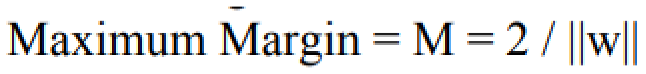
Principal Component Analysis (PCA) is a method employed in exploratory data analysis to reduce the dimensionality of a dataset while preserving maximal variability. This process entails identifying linear combinations of the original variables, continually maximizing variance, and ensuring uncorrelation. These resulting linear combinations are known as principal components (PCs), derived by solving an eigenvalue problem. The PC associated with the largest eigenvalue and its corresponding eigenvector capture the highest variance in the data. In practical applications, PCA is implemented on datasets represented as matrices, where each column corresponds to a variable, and each row corresponds to an observation.

The adaptability of PCs is crucial, as they depend on the specific characteristics of the dataset. Typically used for descriptive purposes rather than inferential ones, PCA can be based on either covariance or correlation matrices. PCs offer an optimal representation of the data, and their variances serve as a key metric for evaluating the effectiveness of dimensionality reduction.The proportion of total variance explained by each PC is an important measure, and users often decide how many PCs to retain based on a predetermined percentage of the total variance. In terms of application , PCA has broad applications in many different fields and is especially useful for displaying high-dimensional data in two or three dimensions as well as outlier detection .It provides the ability to simplify complex data sets while retaining essential information, making it a widely used technique in data analysis.

6.1.7 Support Vector Machine

Support vector machines (SVMs) have emerged as an attractive solution for binary classification tasks, especially when multiple hyperplanes can classify the data. The goal turns to determining the hyperplane with maximum separability, introducing the concept of maximum margin classification.

This approach seeks to maximize the distance between the hyperplane and the closest points on the convex hull of the two data sets, represented by maximizing the amplitude.The maximum amplitude is expressed as



$$\text{Maximum Margin} = M = 2 / ||w||$$

where w is the weight vector.

The solution includes solving the quadratic optimization problem with linear constraints, introducing Lagrange multipliers (αi) and formulating the dual problem.

The SVM classifier integrates essential variables such as a bounding variable (ξi) to measure error and a regularization parameter (C) to control the trade-off between achieving a reasonable level of training.

In addition to the linear model, kernels play a central role in SVM, serving as the fundamental tool for processing nonlinear data patterns. Through the use of a kernel, SVM can map data into a higher dimensional space, making it linearly separable. This transformation, called the kernel trick, involves representing the SVM algorithm using inner products, performing a non-linear mapping of the original data to higher dimensions, and using the function of the kernel to represent the dot product without the need for explicit computation in the feature space. Polynomial kernels stand out as a commonly used choice for nonlinear modeling, with quadratic polynomials often preferred to address specific challenges.

In conclusion, SVM has proved to be a powerful tool for approximating and generalizing training data; however, its effectiveness is closely related to the complexity introduced by the chosen kernel.The tuning parameters is not inherently guided or determined by the SVM model itself and needs to be experimented with different parameter values and techniques like cross-validation to find the configuration that best suits the specific characteristics of the data.

6.1.8 Logistic Regression

Logistic regression models rely on assessing the probability of a binary outcome, where one outcome is designated as the event of interest. Logistic regression employs the natural logarithm of these odds as a regression function. For a single predictor

X, the model takes the form

$$\ln[\text{odds}(Y = 1)] = \beta_0 + \beta_1 X$$

where where

Y is the outcome,B0 is the intercept, and B1 is the regression coefficient.

The odds ratio , represents the probability ratio corresponding to a variation of a unit . Although the odds ratio is different from relative risk, it serves as an approximation when the probability of an event occurring is low.

Logistic regression primarily reports odds ratios due to their natural origin from the model. The main interpretation of odds is 100 × (odds − 1) 100 × (odds − 1)

gives the percentage change in the odds of the event for one unit increases from X.
A positive value represents an increase, while a negative value represents a decrease.
The assumption of linearity between X and the logarithm of the probabilities is the basis for this interpretation.
Such methods as Goodness-of-fit measures, including R^2 and c-statistics, evaluate the overall performance of the model.

6.1.9 Ridge Regression and Lasso Regression

Regression algorithms like Ridge and Lasso use a penalty term to combat overfitting. That can be called the regularization term. It works especially well when there is multicollinearity, or a high degree of correlation between independent variables. Regularization keeps the model from overfitting data points by purposefully adding more errors to it, guarantees consistent performance with both training and testing datasets, even if this could lead to relatively worse performance with training datasets.

6.1.9.1.Ridge Regression

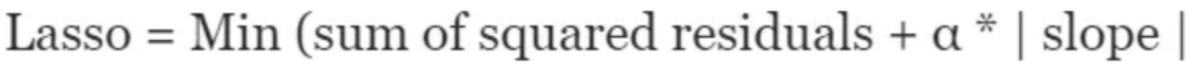
Ridge regression vuses the square of the slope rather than the absolute value as a bias term to prevent overfitting. The penalty term can be expressed mathematically as

$$\text{Ridge} = \text{Min (sum of squared residuals} + \alpha * \text{slope}^2)$$

In order to avoid overfitting, both algorithms seek to increase bias and decrease variance. The way they handle coefficients is where they diverge most. While Lasso can shrink coefficients all the way to zero, Ridge tries to decrease coefficients closer to zero without really reaching zero. Furthermore, Lasso makes feature selection easier by eliminating superfluous characteristics, which makes it appropriate for datasets with fewer features; in contrast, Ridge works better with smaller datasets.

6.1.9.2.Lasso Regression

Lasso Regression uses the absolute value of the magnitude of the coefficient. Mathematically, It is expressed as:



$$\text{Lasso} = \text{Min (sum of squared residuals} + \alpha * | \text{slope} |$$

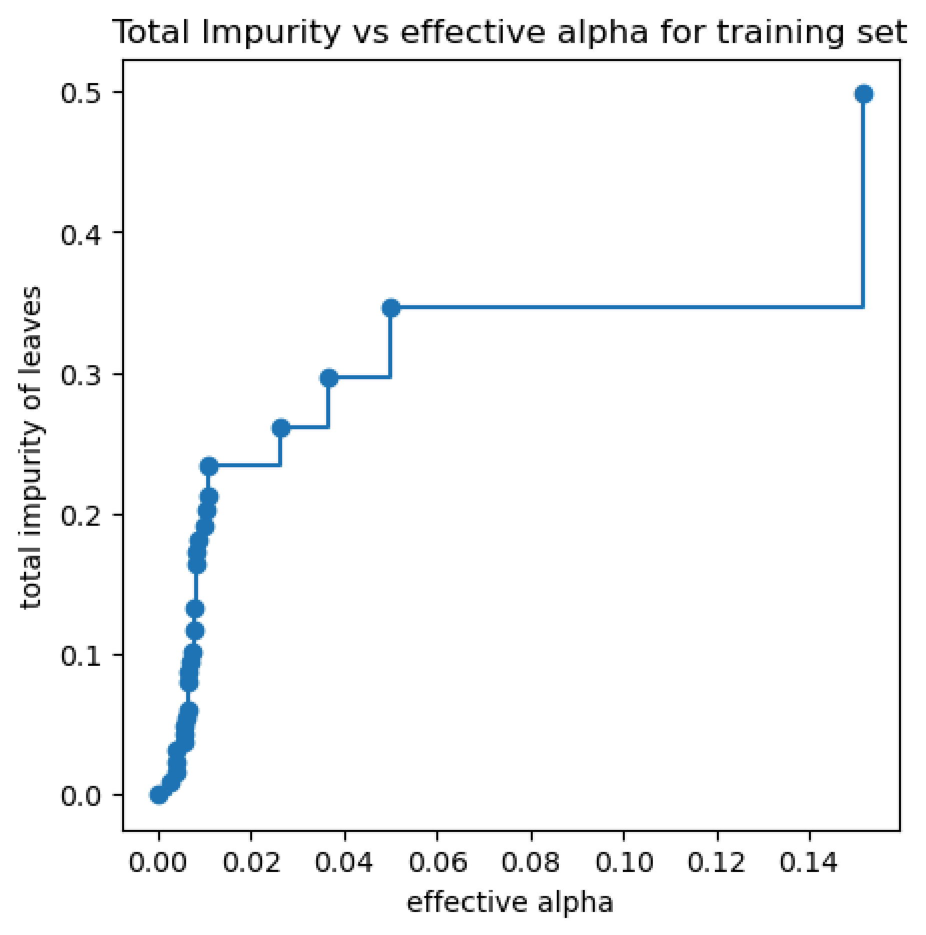
where the term α * | slope | is the penalty term.

6.2. Results Analysis and Interpretation

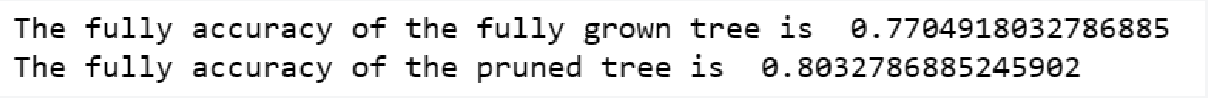
6.2.1.Decision Trees

The maximum depth of the fully grown tree is 9. After training on the preprocessed data, the classification_tree had a depth of 7 as chosen for checking , 36 leaves, and achieved an impressive accuracy of 97.9% on the training set. Following this, we made predictions on the test set, resulting in an overall accuracy of approximately 77% .

Pruning plays a crucial role in decision trees to address the problem of overfitting. By constraining the tree's size, pruning prevents it from memorizing irrelevant details in the training data, thereby promoting improved generalization and robustness. Moreover, pruning simplifies the tree's structure, enhancing its interpretability and resource efficiency. This process ensures a balanced trade-off between model complexity and accuracy, resulting in decision trees that capture essential patterns without unnecessary complexities. During the pursuit of achieving the best possible pruning for the Decision Tree The graph below suggests that a balance between trade-off and total impurity was achieved with an effective alpha of roughly 0.05.



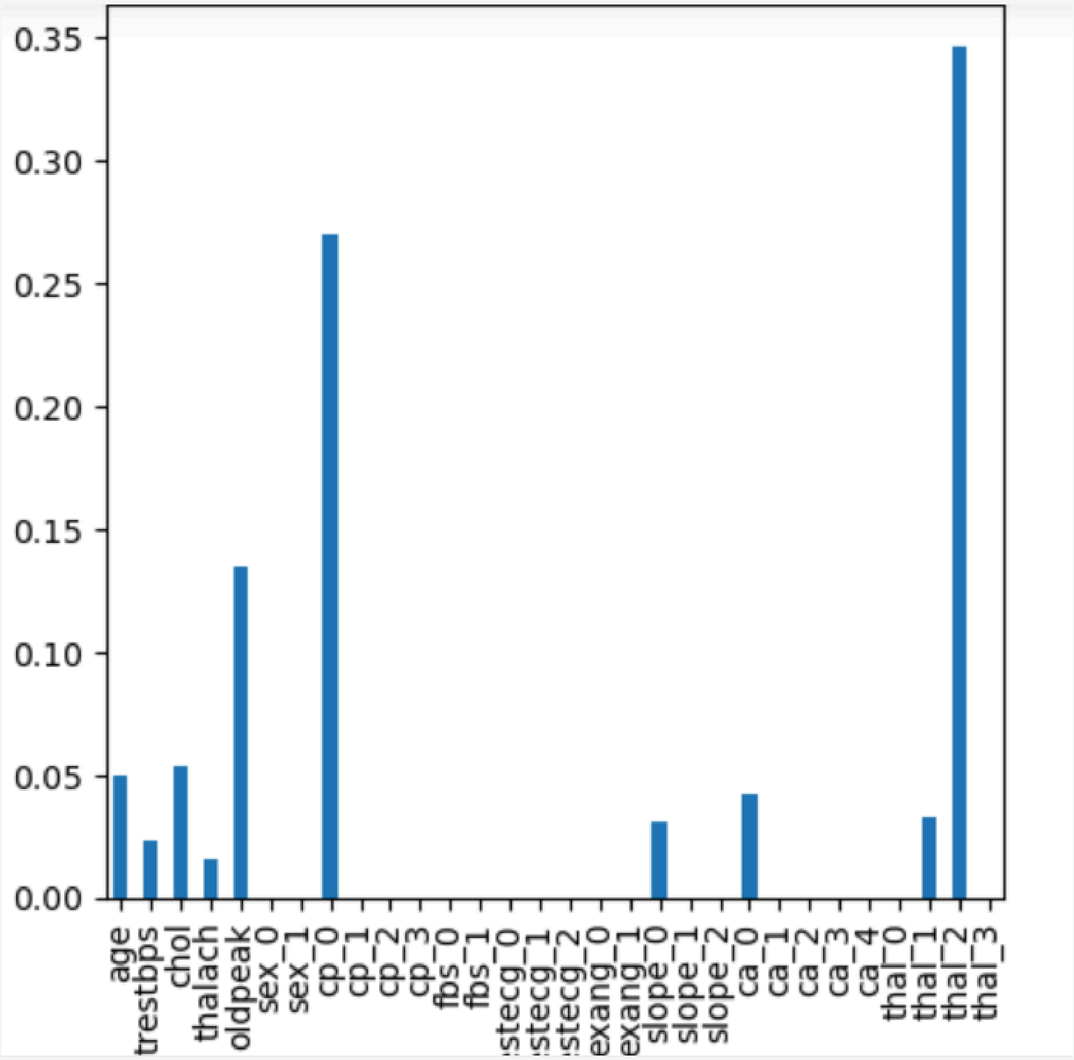
To further refine the model, a comprehensive grid search was conducted over combinations of max_depth and class_weight. The best estimator identified from this search had a max_depth of 4 and a class_weight of {0: 1, 1: 3}, demonstrating the effective pruning of the Decision Tree. Finally, the fully grown and pruned trees were evaluated on the test set. The accuracy of the fully grown tree was 78.69%, while the pruned tree achieved an accuracy of 77.05%, indicating that pruning significantly compromised predictive performance.



```
The fully accuracy of the fully grown tree is  0.7704918032786885
The fully accuracy of the pruned tree is  0.8032786885245902
```

Additionally, the root mean squared error (RMSE) between the predicted and actual outcomes was found to be 0.4791.

The feature importance ranking graph highlights the significant role played by the variables oldpeak, cp_0, and thal_2 in predicting heart attacks.



In conclusion, the optimization process, which considered both cost-complexity pruning and maximal depth selection, resulted in a well-tailored Decision Tree Classifier that strikes a balance between model complexity and predictive accuracy.

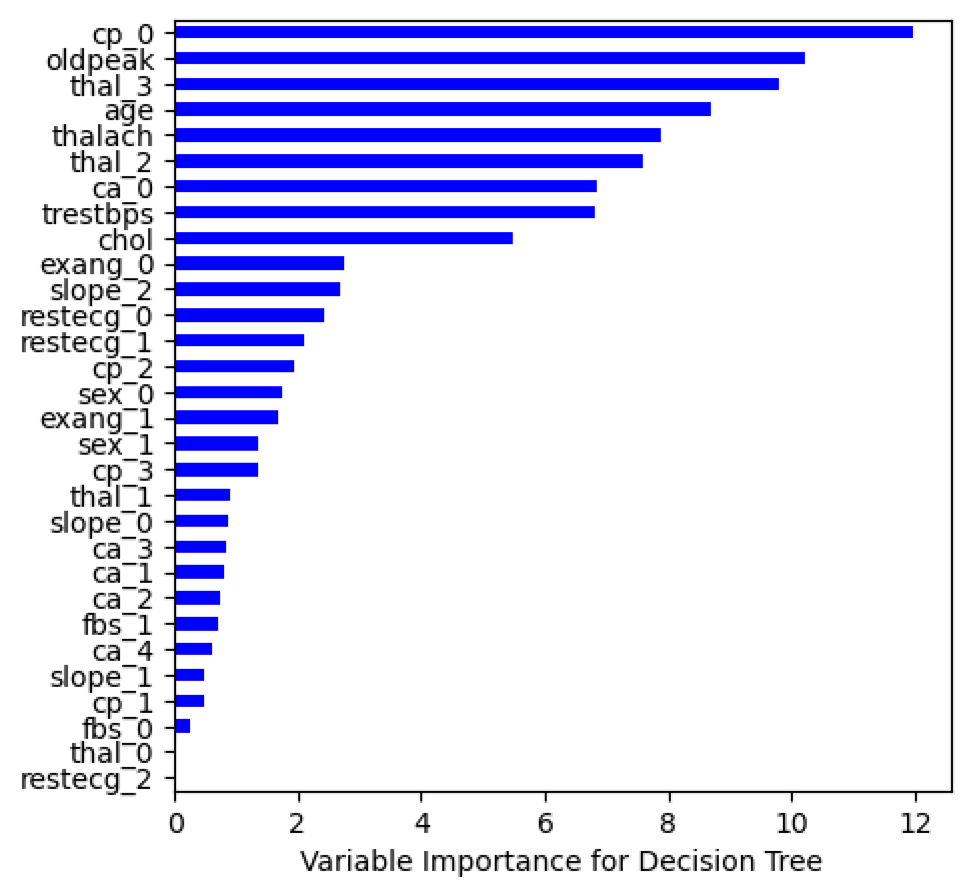
Image for **final optimal Decision tree** has been attached as Image 3 of Appendix

6.2.2.Random forest and Bagging

Random Forest and bagging methods outperform standalone decision trees by addressing the issue of overfitting and improving predictive accuracy. Theoretically, Random Forest builds an ensemble of trees, with each tree trained on a random subset of the data while bagging, a more comprehensive technique, involves training models on diverse subsets of the data.These approaches helps mitigate overfitting, reducing variance, improving stability and enhances performance compared to individual decision tree.

The Random Forest model was created using nine decision trees, with an emphasis on diversity by utilizing random subsets of the training data. On the other hand, the Bagging approach employed a total of 30 decision trees, each trained independently on different subsets of the data. Both the Random Forest and Bagging ensemble methods showed significant improvements compared to the standalone Decision Tree model in terms of accuracy and predictive performance. The Decision Tree achieved an accuracy of approximately 77.05%, while both the Random Forest and Bagging models achieved a higher accuracy of 78.69%. This improvement suggests that the ensemble techniques effectively addressed the overfitting issues associated with the standalone Decision Tree, allowing for better generalization to new, unseen data.

Moreover, the importance of features in Random Forest has been altered, with thal_3 now being deemed more important than thal_2 compared to Decision Tree.



Additionally, the root mean squared error (RMSE) serves as a useful metric to evaluate the predictive accuracy of the models. The Decision Tree had an RMSE of 0.4791, while both the Random Forest and Bagging models had a lower RMSE of 0.4616. The reduced RMSE values indicate that the ensemble methods produced more accurate predictions, highlighting their effectiveness in minimizing prediction errors. When comparing the Random Forest and Bagging models, their accuracies and RMSE values were nearly identical. However, the Bagging approach demonstrated a slight advantage with a slightly lower RMSE. This suggests that the ensemble of decision trees in Bagging collectively contributed to a more refined and accurate predictive model.

Basically, both the Random Forest and Bagging ensemble methods outperformed the standalone Decision Tree in terms of accuracy and predictive accuracy. While the Random Forest and Bagging models showed similar performance, the slight advantage observed with Bagging emphasizes the importance of leveraging multiple decision trees to enhance model robustness and accuracy in the context of heart attack prediction in this analysis.

6.2.3.Boosting

Boosting is employed to enhance the accuracy of decision trees which can be achieved by training several weak learners in sequence, usually shallow decision trees. Each subsequent tree is designed to correct the errors of the ensemble, leading to a more robust and precise predictive model. The Boosting model was configured with 500 decision trees, a learning rate of 0.01, and a maximum depth of 4. To optimize the number of trees and prevent overfitting, early stopping was implemented, resulting in the determination of an optimal number of trees as 185.More in detail, the Gradient Boosting algorithm utilizes early stopping to halt training when the model's performance on a validation set reaches a plateau. This technique helps to select the optimal iteration for a more balanced and generalizable model. The staged_predict method is used to find this iteration, preventing overfitting .

Compared to previous ensemble methods and the standalone Decision Tree, the Boosting model exhibited notable improvement with an accuracy of 80.33%. It outperformed the Random Forest (78.69%), Bagging (78.69%), and Decision Tree (77.05%) models. The mean squared error (MSE) also decreased to 0.4339, indicating a more accurate predictive performance.

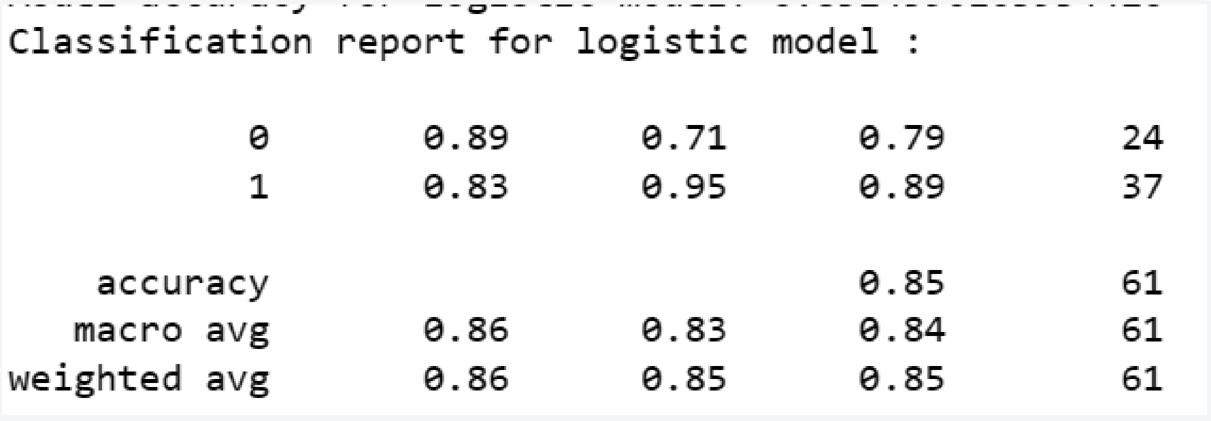
In conclusion, the Gradient Boosting model demonstrated significant improvement in accuracy and predictive accuracy compared to Decision Trees, Bagging, and Random Forest. The iterative refinement process and careful selection of the optimal number of trees contributed to the success of the Boosting approach in the context of heart attack classification.

Feature important ranking for Boosting has been attached as Image 4 of Appendix

6.2.4.Logistic regression

The logistic regression model's confusion matrix provides a detailed evaluation of its classification performance in predicting heart attacks. The model accurately identified 35 instances with a heart attack (True Positives) and 17 instances as not having a heart attack (True Negatives). However, it also made 7 false positive classifications and missed 2 instances with a heart attack (False Negatives). The model's overall accuracy of 85.25% demonstrates its effectiveness in making correct predictions.

The following table displays the report on classification:



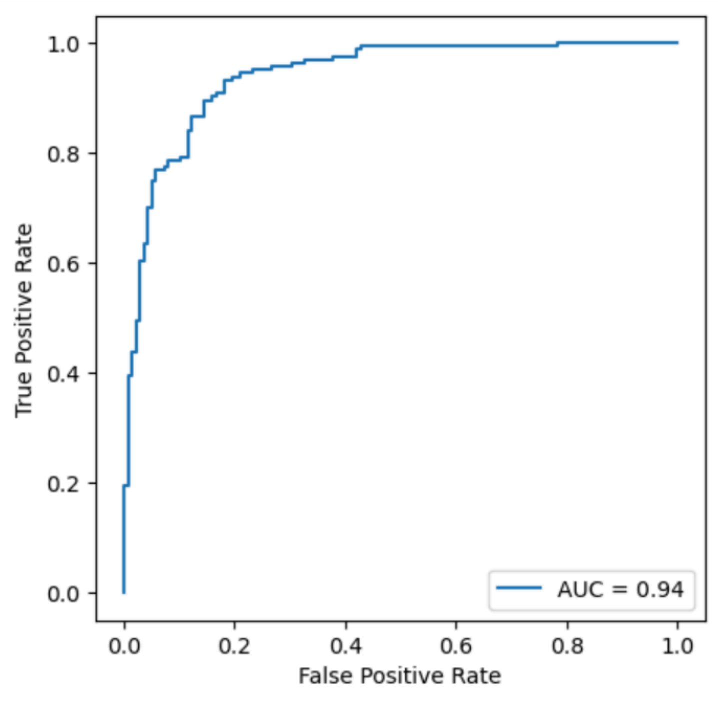
```
Classification report for logistic model :

           0       0.89      0.71      0.79        24
           1       0.83      0.95      0.89        37

    accuracy                           0.85        61
   macro avg       0.86      0.83      0.84        61
weighted avg       0.86      0.85      0.85        61
```

When the heart attack prediction model identifies an instance as positive, it is correct around 83.33% of the time, indicating a precision of 83.33%. This means that out of all the instances predicted as having a heart attack, approximately 83.33% are accurate predictions, while the remaining 16.67% are false alarms. This logistic model has the ability to minimize the rate of false positives is crucial in applications where accurate identification is paramount to avoid unnecessary interventions or treatments. On the other hand, a recall of 94.59% means that the heart attack prediction model successfully captures and identifies around 94.59% of all actual instances of heart attacks in the dataset. This high recall is indicative of the model's ability to minimize false negatives, ensuring a comprehensive coverage of true positive predictions.In the given analysis, it is imperative to prioritize a high recall rate in order to reduce the occurrences where the model fails to detect a genuine heart attack. This will greatly benefit healthcare applications.

ROC curve is also one of the methods for model's performance assessment. The ROC curve shown below demonstrates a strong ability to distinguish between positive and negative instances, with an AUC of 0.94. This high discriminative power highlights the effectiveness of the model in identifying individuals with and without heart attacks.

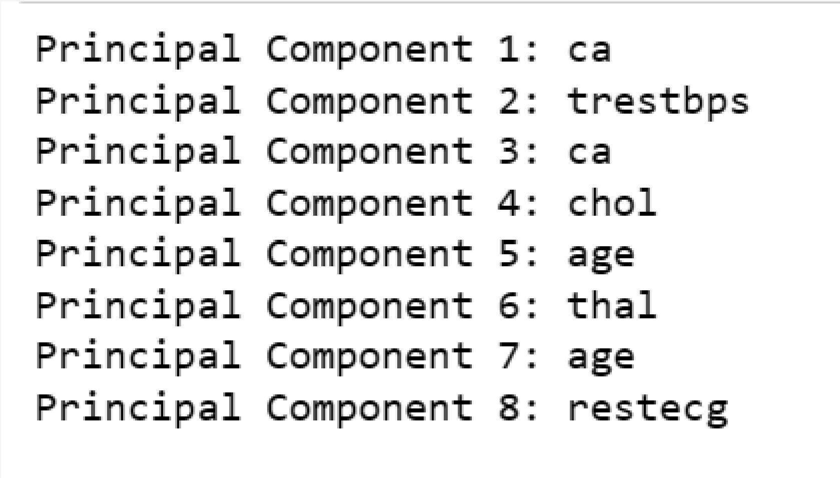


The steep rise in the curve indicates effective discrimination without compromising specificity. Moreover, the AUC assists in selecting an optimal probability threshold for predictions. For this reason, the AUC of 0.94 emphasizes the model's reliability in accurately identifying individuals at risk of heart attacks.

6.2.5.Logistic Regression Modeling After Principal Component Analysis (PCA)

To address concerns about multicollinearity that was detected by VIF , Principal Component Analysis (PCA) was utilized to reduce dimensionality while preserving important information. The eight selected principal components,

labeled 'PC1' through 'PC8,' collectively accounted for 86.5% of the variance in the dataset. The variance explained by each component is as follows: 'PC1' (21.31%), 'PC2' (11.86%), 'PC3' (9.36%), 'PC4' (9.12%), 'PC5' (7.87%), 'PC6' (7.46%), 'PC7' (6.65%), and 'PC8' (5.99%). The decision to retain these eight components was based on the observation that increasing the number of components did not significantly contribute to the total explained variance.
Upon examining the loadings of these components, it became clear that certain variables, such as 'ca' (number of major vessels colored by fluoroscopy), 'trestbps' (resting blood pressure), 'chol' (serum cholesterol), 'thalach' (maximum heart rate achieved), and 'age,' played crucial roles in forming the principal components.



```
Principal Component 1: ca
Principal Component 2: trestbps
Principal Component 3: ca
Principal Component 4: chol
Principal Component 5: age
Principal Component 6: thal
Principal Component 7: age
Principal Component 8: restecg
```

Despite this careful feature selection process, the resulting logistic regression model trained on the reduced feature set exhibited a lower test accuracy of 72.1%. Although PCA tackles multicollinearity, the observed decline in performance could be ascribed to the potential oversimplification of non-linear associations and the omission of crucial information during dimensionality reduction. This suggests that the influence of multicollinearity may not be significant in this particular scenario.
Upon further analysis of the variance accumulation and the loadings of the principal components, it becomes evident that a well-balanced approach to

feature reduction is crucial. This approach ensures that the complexity of the model aligns with the intricacies of the dataset. Nevertheless, despite the fact that eight components capture the maximum variance in the dataset, increasing their number does not result in any significant improvement. Consequently, it suggests that PCA may not provide any additional advantages in this particular scenario.Alternative approaches such as Lasso and Ridge regression will be explored to address feature selection and multicollinearity concerns later on in this analysis.

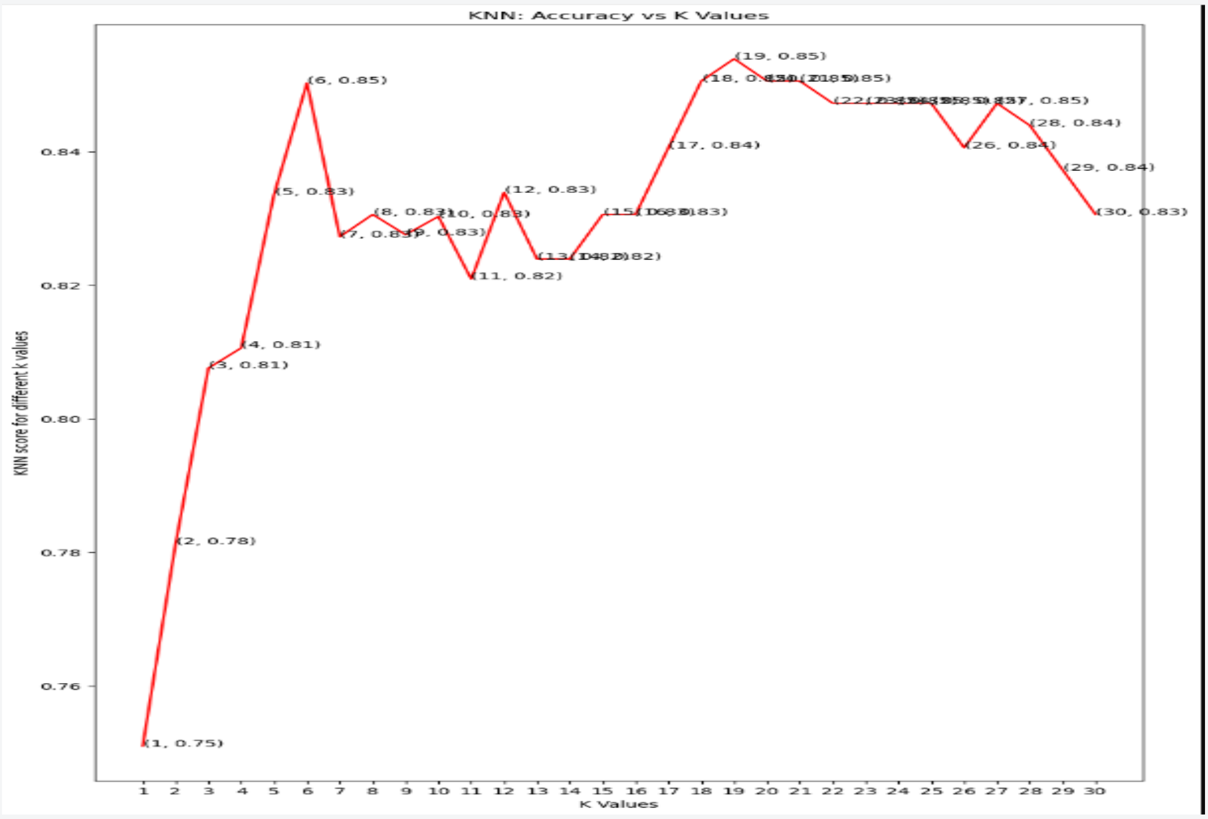
6.2.6.Ridge and Lasso

Addressing multicollinearity is of utmost importance in order to improve the reliability of logistic regression models. To tackle this issue, I have implemented Ridge and Lasso regression techniques as proactive measures for both model selection and feature reduction, addressing multicollinearity issues indirectly. The optimization of alpha which is indeed the penalty term is a crucial step in Ridge and Lasso regularization techniques. The RidgeCV method facilitates the identification of the optimal alpha through cross-validation, ensuring a balanced trade-off between model complexity and accuracy. In this case, alpha is found at 0.5336699231206307. Similarly, LassoCV is employed for alpha selection, considering a range of values and using cross-validation to identify the most effective regularization parameter. The optimal alphas obtained through these methods contribute to the models' ability to penalize and shrink coefficients appropriately, addressing multicollinearity and preventing overfitting. Moreover, the choice of threshold for classification is also a critical factor for Ridge and Lasso regression. Setting the threshold at 0.5 affects the trade-off between sensitivity and specificity.

The Ridge model achieved an excellent test accuracy of 85.25% which is equal to Logistic regression .It demonstrated a mean squared error of 0.125 on the test set. With a mean squared error of 0.126 on the test set, the Lasso model demonstrated how well it minimizes mistakes. Notably, the test accuracy of the Lasso model was even higher, at 88.52% .It is evident that Lasso's test accuracy has improved. The characteristics of model architecture might provide an explanation.Ridge regression tends to shrink coefficients towards zero without completely eliminating them. This approach helps in reducing multicollinearity but may still retain some less relevant features. On the other hand, Lasso regression has a unique feature selection property that can set certain coefficients exactly to zero, effectively excluding specific features from the model. By doing so, Lasso regression can create a sparser model that focuses on the most influential predictors. This feature selection property of Lasso can contribute to improved generalization and higher accuracy on the test set, especially when irrelevant features are successfully excluded.

6.2.7. KNN

The K-Nearest Neighbors (KNN) algorithm relies on the concept of considering the classes of nearby data points to predict the class of a given instance. One of the critical factors that affect the performance of KNN is the selection of the hyperparameter 'k,' which represents the number of neighbors considered during classification. In this analysis, a range of 'k' values from 1 to 30 was explored, and their corresponding accuracies were evaluated using cross-validation.
The resulting accuracy vs K values graph showed fluctuations in performance, highlighting the importance of selecting an appropriate 'k' value.



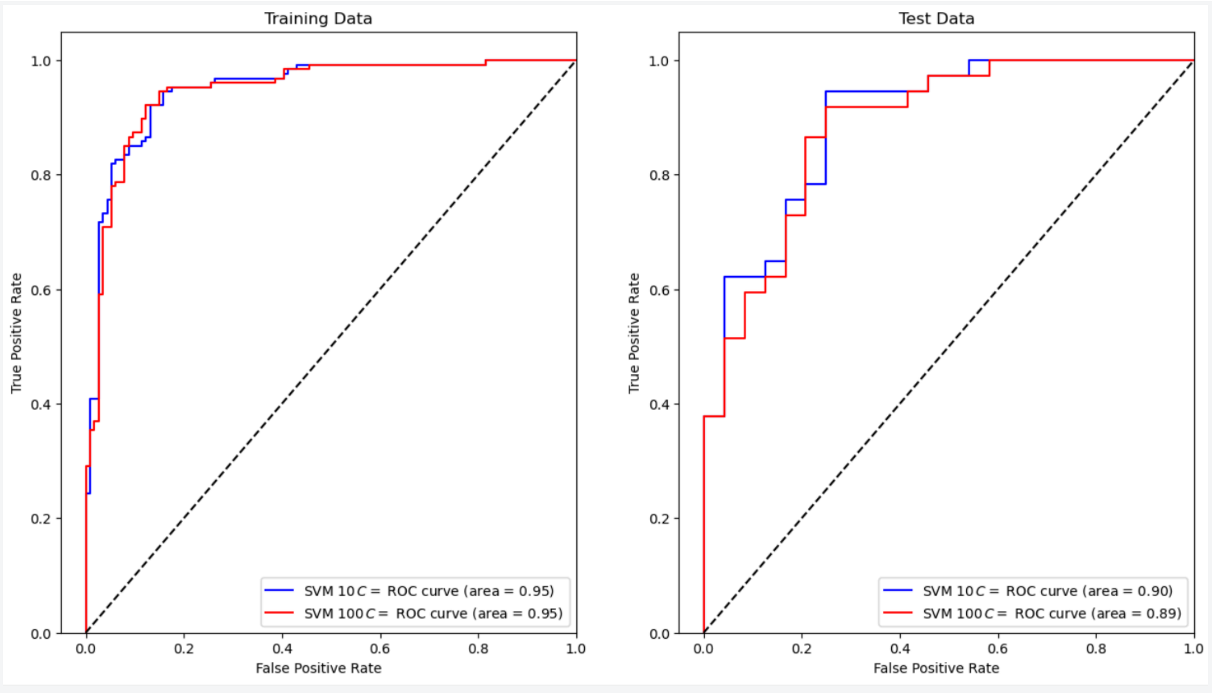
To determine the optimal 'k' value, GridSearchCV was used, and the resulting value of 13 was utilized to train the KNN classifier. The model demonstrated an accuracy of approximately 89% on the test set, indicating the effectiveness of the selected 'k' value in achieving accurate classifications.

Upon examining the graph, it is evident that the optimal value of k, which is 13, did not yield the highest accuracy in the training data.It could be explained by a trade-off between bias and variance. In this case, the chosen 'k' value strikes a balance between capturing underlying patterns and avoiding overfitting. This process underscores the significance of thoughtful hyperparameter tuning in maximizing the predictive capabilities of the KNN algorithm.

6.2.8.Support Vector Machine

We recognized the complexity of the dataset, which may not be visually apparent due to its high-dimensionality. For that reason, we thoroughly examined the various aspects of linear and non-linear kernels to determine the most suitable configuration for our dataset in this extensive analysis of Support Vector Machine (SVM) models for heart attack prediction. Understanding the role of the regularization parameter (C) is crucial in SVM. It allows us to balance the trade-off between achieving a smooth decision boundary and accurately classifying training points. The grid search conducted in this analysis demonstrated the importance of fine-tuning this parameter to achieve optimal model performance.

Our initial linear SVM model achieved an impressive accuracy of approximately 81.97% on the test set, due to a fixed regularization parameter (C=1).At this stage , C parameter is randomly chosen.Posteriorly, we conducted an extensive grid search to optimize the linear SVM model by exploring different regularization parameters (C). After careful evaluation, we discovered that C=0.01 was the optimal value, resulting in a new linear SVM model with a test accuracy of 78.69%. As observed in the ROC curve graph below, increasing the regularization parameter (C) from 10 to 100 results in the lower AUC 0.89 compared to AUC 0.90 for the value of C = 10.The blue line in the test data graph exhibits numerous points where it achieves a higher True Positive Rate while maintaining the same False Positive Rate as the red line. It can be inferred that an increased value of C might result in a slightly less efficient model when it comes to differentiating between positive and negative instances.



In terms of accuracy,there is a decrease of 3.28% in test accuracy when increasing the value of C . This is because a higher value(C = 100) can cause overfitting by closely fitting the training data,capturing noise and specificities in the data which can lead to poor generalization to new unseen data.

Moving forward, we delved into non-linear SVM kernels, specifically radial basis function (RBF) kernels, as we acknowledged the potential complexity of underlying patterns in the high-dimensional feature space. Through another grid search, this time considering both C and gamma hyperparameters, we identified optimal values of C=1 and gamma=0.1 for the RBF kernel SVM model. This model achieved a test accuracy of 77.05%

The RBF kernel is known for its enhanced flexibility, allowing it to capture non-linear connections effectively. However, in the case of a limited dataset with only 302 observations, the RBF kernel may be prone to overfitting, resulting in lower test accuracy compared to the linear model. Hence, it is advisable to opt for the optimal linear SVM model in this scenario.

7.Conclusion and Recommendations

7.1. Summary of Findings:

During our extensive analysis of machine learning models for predicting heart attacks, we discovered that each model possessed unique characteristics, which necessitated a thorough evaluation of their strengths and limitations.

7.1.1 Logistic Regression:

In our exploration of machine learning models for predicting heart attacks, Logistic Regression emerged as a valuable tool, striking a balance between interpretability and accuracy. With an impressive accuracy rate of 72.13%, Logistic Regression provided clear insights into the importance of different features, enabling a straightforward understanding of its decision-making process.

7.1.2 Decision Tree:

The Decision Tree model showcased robust performance in predicting heart attacks, achieving an accuracy rate of 80.33%. Its strength lies in its ability to handle feature interactions and visually represent decision-making. However, Decision Trees are susceptible to overfitting, particularly in the presence of outliers, and their ability to capture complex relationships may be limited.

7.1.3 Lasso Regression:

Among the models we explored, Lasso Regression emerged as the top performer, boasting an impressive accuracy rate of 88.52%. In addition to its high predictive accuracy, Lasso Regression also possesses built-in feature selection capabilities, making it a powerful tool for extracting relevant features in datasets with high dimensions. Despite its success, it is important to consider potential performance degradation when dealing with certain scenarios.

7.1.4 Random Forest:

Random Forest, a model that combines multiple Decision Trees, demonstrated competitive accuracy, reaching 78.69%. It addressed some of the limitations of individual Decision Trees by mitigating overfitting through ensemble techniques. However, it is worth noting that Random Forest comes with computational demands and may lack the interpretability offered by individual trees.

7.1.5 Support Vector Machine

SVM has demonstrated its effectiveness in dealing with high-dimensional data and capturing complex relationships by utilizing various kernels. It has achieved an accuracy of 77.05%, showcasing its potential. However, the computational requirements of SVM may limit its usage in large-scale applications, and careful tuning of hyperparameters is crucial for optimal performance.

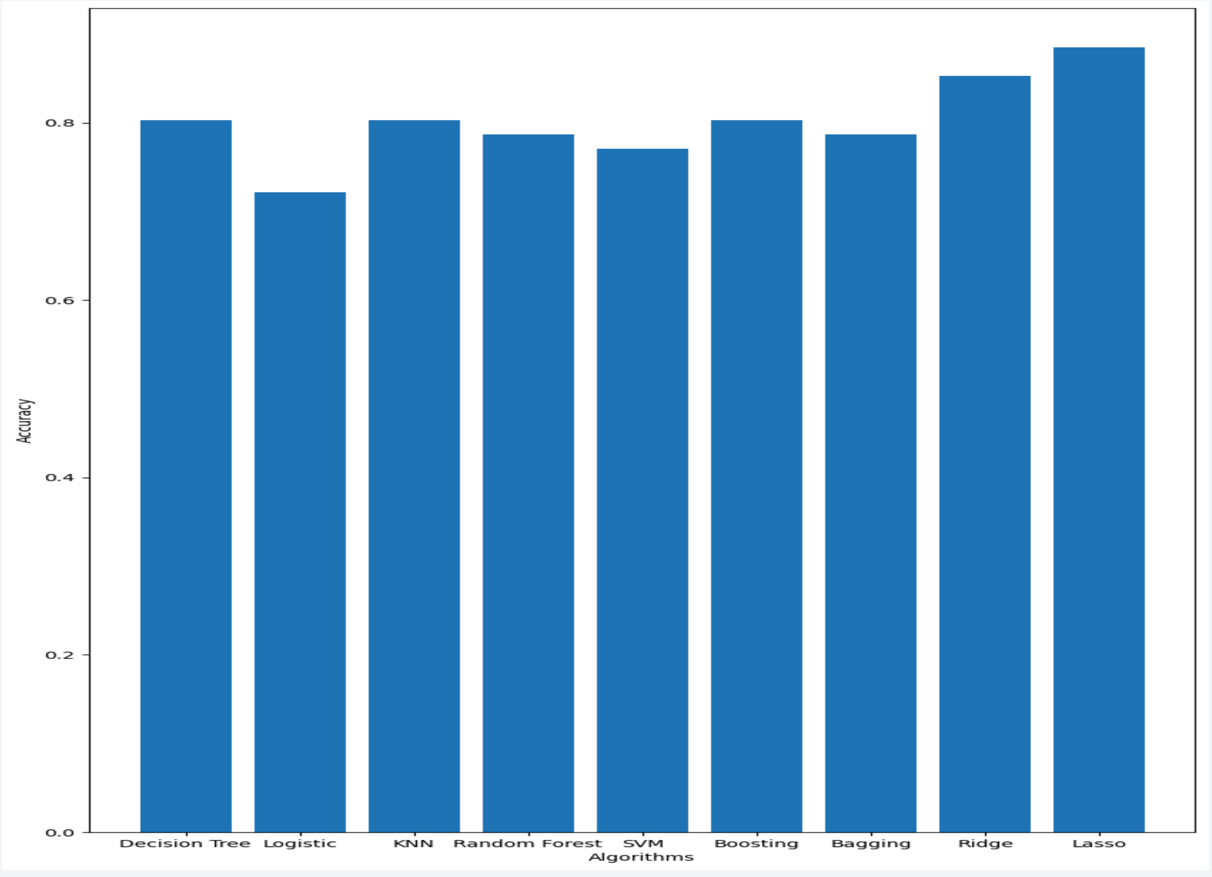
7.1.6 Boosting and Bagging

Boosting and Bagging Classifier, have shown competitive accuracies of 80.33% and 78.69% respectively. These methods excel in enhancing model performance by combining weaker learners. While boosting and bagging can improve accuracy, they may also come with increased computational demands.

7.2 Comparative Evaluation and Model selection recommendation.

A comparative analysis of the models has revealed distinct trade-offs between accuracy, interpretability, and computational efficiency.

Lasso regression has emerged as the top performer, striking a balance between high accuracy and feature selection and is recommended for heart attack prediction in this analysis based on the highest accuracy value shown below:



Random Forest provides a compromise between accuracy and complexity while Logistic Regression and Decision Tree are recommended when interpretability is of utmost importance.To clarify, Logistic Regression and Decision Trees are ideal for situations that prioritize interpretability. These models offer straightforward and comprehensible insights into the impact of individual heath features on heart attack predictions.The Decision tree approach highlights typical angina chest pain and genetic blood disorders as the top two crucial features for predicting heart conditions. This attribute proves to be particularly advantageous in situations where conveying outcomes to non-technical parties, except for nurses or doctors, or complying with regulatory standards that mandate a clear-cut methodology is crucial.

The current analysis has not explicitly considered metrics beyond accuracy such as precision and recall. However, this exploration would serve as a valuable avenue for gaining deeper insights and a more nuanced understanding of model performance in the given domain. For example, in medical diagnoses like predicting heart attacks, the focus might be on minimizing false negatives to ensure that potential cases are not overlooked. In such cases, giving high importance to recall becomes crucial, even if it results in more false positives. On the other hand, if the objective is to minimize unnecessary interventions and concentrate on accurately confirmed cases, precision becomes more important. Therefore, the choice between precision and recall should be guided by the overarching goal, ensuring that the selected model aligns with the intended purpose and provides outcomes that are in line with the desired priorities.

7.3.Prospective Directions

Exploring cutting-edge techniques like ensemble methods, model stacking, and diverse feature engineering approaches shows great potential in enhancing predictive abilities, thereby driving the advancement of predictive capabilities in the healthcare system for heart disease. Additionally, it is crucial to highlight the significance of regularly updating and recalibrating the model to adapt to the ever-changing healthcare data landscape, ensuring consistent accuracy in predicting heart-related events. This not only emphasizes the technical aspects but also demonstrates a broader dedication to continuously improving predictive models for the betterment of healthcare outcomes.

Reference:

1. Wichmann, N. (2021, September 22). Retrieved November 23, 2023, from https://www.nature.com/articles/s41598-020-72685-1?fbclid=IwAR3H2kxqECoXCrAet6I5Ap5idByjG7thlnko1jQ4zTlA7-5Q0gXYwuYWKl8
2. Ramalingam, V. (2019, August 5). *(PDF) Heart disease prediction using machine learning techniques: A survey*. ResearchGate. Retrieved November 23, 2023, from https://www.researchgate.net/publication/325116774_Heart_disease_prediction_using_machine_learning_techniques_A_survey
3. Sharma, A. (2016, 05 14). *(PDF) Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review*. ResearchGate. Retrieved November 23, 2023, from https://www.researchgate.net/publication/303318878_Theoretical_Study_of_Decision_Tree_Algorithms_to_Identify_Pivotal_Factors_for_Performance_Improvement_A_Review
4. Bardan Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2023, June 16). *An Evaluation of Bagging, Boosting, and Random Forests for Land-Cover Classification in Cape Cod, Massachusetts, USA*. YouTube. Retrieved November 23,

2023, from https://www.tandfonline.com/doi/epdf/10.2747/1548-1603.49.5.623?needAccess=true

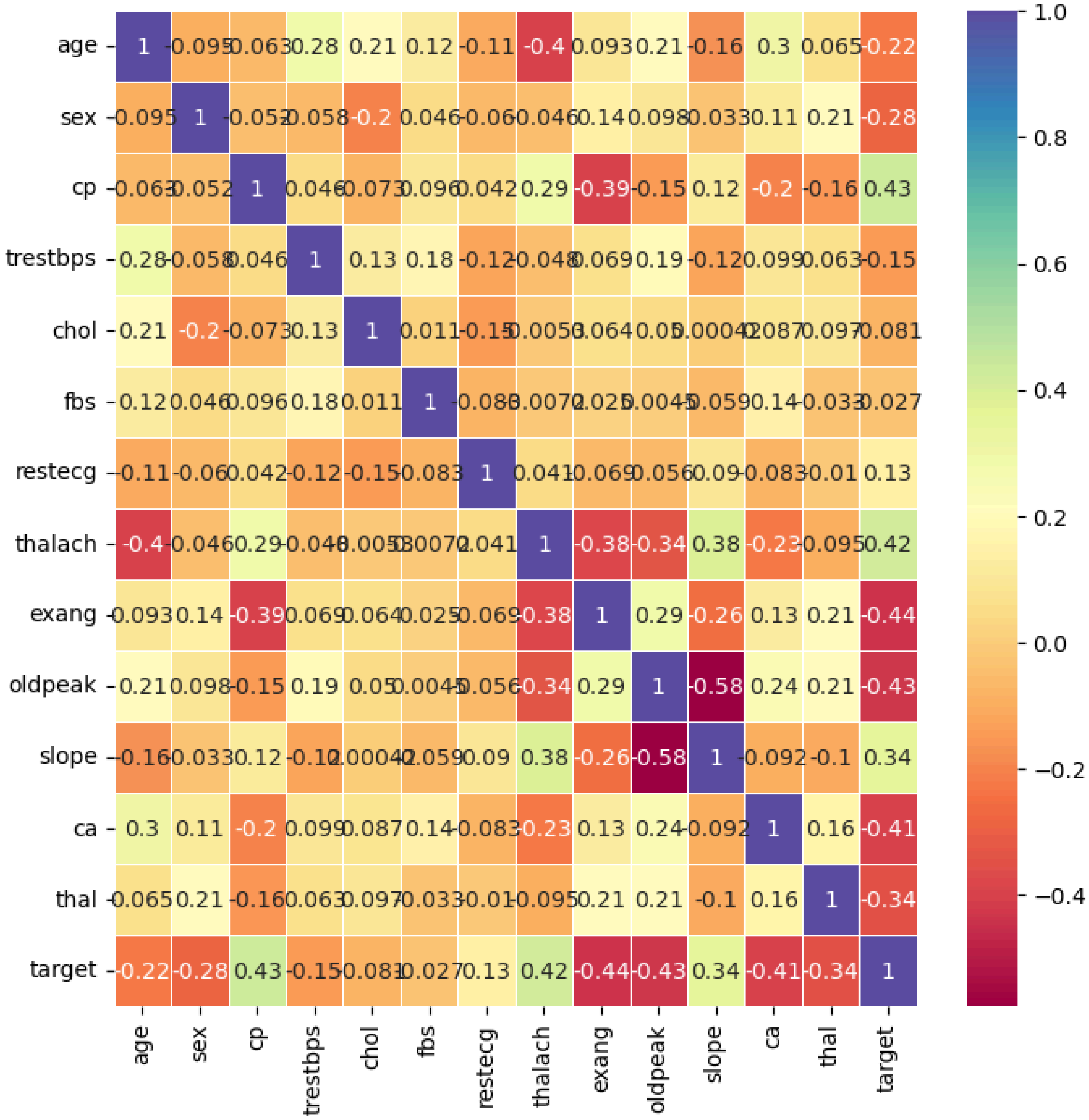
5. Wichmann, N. (2021, September 22). *Introduction to machine learning: k-nearest neighbors*. Retrieved November 23, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/?fbclid=IwAR0n_OOjp4kUSSHsb1uFDXBF13flh7dmyDikjgzn_KHAEq_73kbyW9YzUJs#sec-a.t.btitle
6. Pandey, A., & Achin Jain, A. (2023, June 16). *Comparative Analysis of KNN Algorithm using Various Normalization Techniques*. Retrieved November 23, 2023, from https://mecs-press.net/ijcnis/ijcnis-v9-n11/IJCNIS-V9-N11-4.pdf?fbclid=IwAR0Fyz5h8mfOolphqYbg228a3Wgr0k8_fG9TRCG-7UEce25aYNcR0S0OqGk
7. Wichmann, N. (2021, September 22). *Principal component analysis: a review and recent developments*. Retrieved November 23, 2023, from https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202?fbclid=IwAR2Hoz0qz1MmiY008PZLN6AlbvP6UWyTEYZLgFbDuLI63oZLc57SSjyyBNA
8. Jakkula, V. (2023, June 16). *Tutorial on Support Vector Machine (SVM)*. Retrieved November 23, 2023, from https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf?fbclid=IwAR0yLSdk4bbdRt0t2we80I2Lot9WiOLp4PgWO0vDOWJ0J4kg1zQidZocsKI
9. Parida, J. (2022, September 12). *Understanding Ridge and Lasso Regression | by Jayanta Parida*. Retrieved November 23, 2023, from https://medium.com/@jayantspeaks/understanding-ridge-and-lasso-regression-3405dcada1a4
10. Zach, Z. (2023, June 16). *How to Read a Correlation Matrix*. Retrieved November 23, 2023, from https://www.statology.org/how-to-read-a-correlation-matrix/?fbclid=IwAR2eAODo1NsgWqQ2RV6K5cWEIAZGYevE0bq-9RI4zcdxUfptMFXisFOcc8c

11. Zach, Z. (2023, June 16). *A Guide to Multicollinearity & VIF in Regression*. Retrieved November 23, 2023, from https://www.statology.org/multicollinearity-regression/?fbclid=IwAR0YOa-kzde39pEvpeOWCDw4jB7eLxa5m6c4e9FGfffCaCF3ML2HWY8QdT4

12.

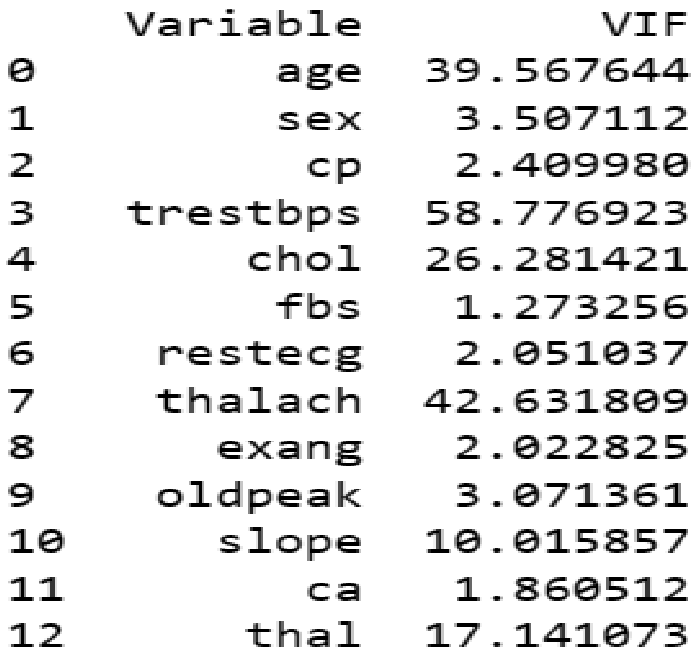
12. Appendix:

Image 1: Correlation heatmap



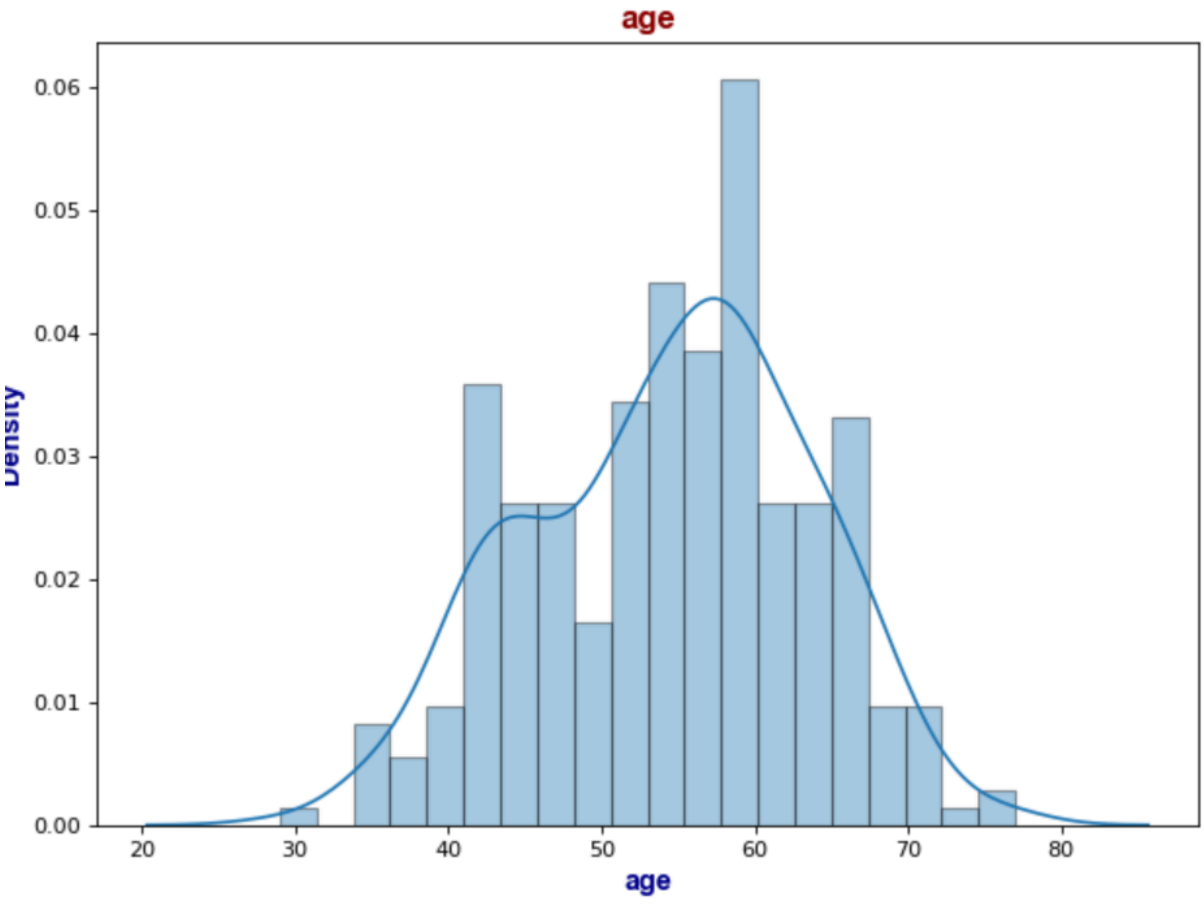
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.095	-0.063	0.28	0.21	0.12	-0.11	-0.4	0.093	0.21	-0.16	0.3	0.065	-0.22
sex	-0.095	1	-0.052	-0.058	-0.2	0.046	-0.06	-0.046	0.14	0.098	-0.033	0.11	0.21	-0.28
cp	-0.063	-0.052	1	0.046	-0.073	0.096	0.042	0.29	-0.39	-0.15	0.12	-0.2	-0.16	0.43
trestbps	0.28	-0.058	0.046	1	0.13	0.18	-0.12	-0.048	0.069	0.19	-0.12	0.099	0.063	-0.15
chol	0.21	-0.2	-0.073	0.13	1	0.011	-0.15	-0.0053	0.064	0.05	-0.00042	0.087	0.097	-0.081
fbs	0.12	0.046	0.096	0.18	0.011	1	-0.083	-0.0072	0.025	0.0045	-0.059	0.14	-0.033	-0.027
restecg	-0.11	-0.06	0.042	-0.12	-0.15	-0.083	1	0.041	-0.069	-0.056	0.09	-0.083	-0.01	0.13
thalach	-0.4	-0.046	0.29	-0.048	-0.0053	-0.0072	0.041	1	-0.38	-0.34	0.38	-0.23	-0.095	0.42
exang	0.093	0.14	-0.39	0.069	0.064	0.025	-0.069	-0.38	1	0.29	-0.26	0.13	0.21	-0.44
oldpeak	0.21	0.098	-0.15	0.19	0.05	0.0045	-0.056	-0.34	0.29	1	-0.58	0.24	0.21	-0.43
slope	-0.16	-0.033	0.12	-0.12	-0.00042	-0.059	0.09	0.38	-0.26	-0.58	1	-0.092	-0.1	0.34
ca	0.3	0.11	-0.2	0.099	0.087	0.14	-0.083	-0.23	0.13	0.24	-0.092	1	0.16	-0.41
thal	0.065	0.21	-0.16	0.063	0.097	-0.033	-0.01	-0.095	0.21	0.21	-0.1	0.16	1	-0.34
target	-0.22	-0.28	0.43	-0.15	-0.081	-0.027	0.13	0.42	-0.44	-0.43	0.34	-0.41	-0.34	1

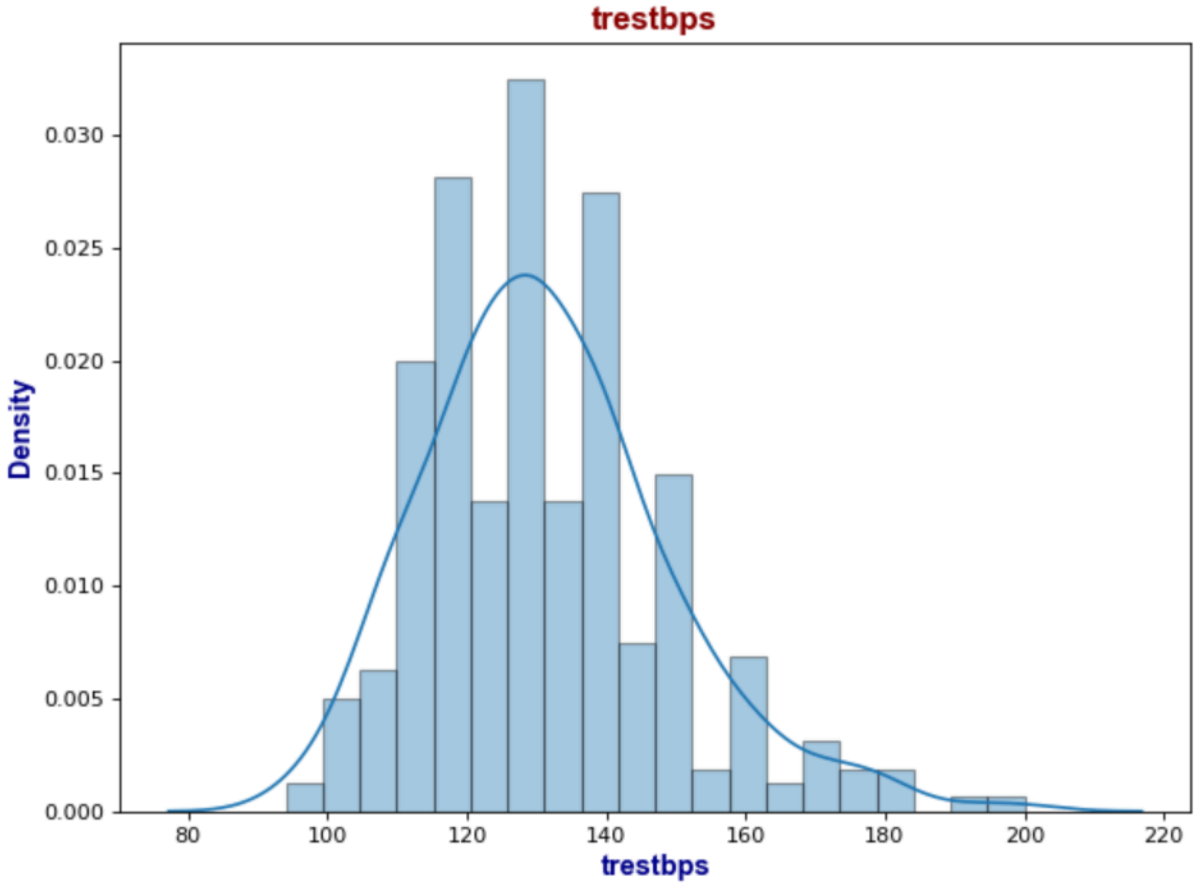
Table 1 : VIF results

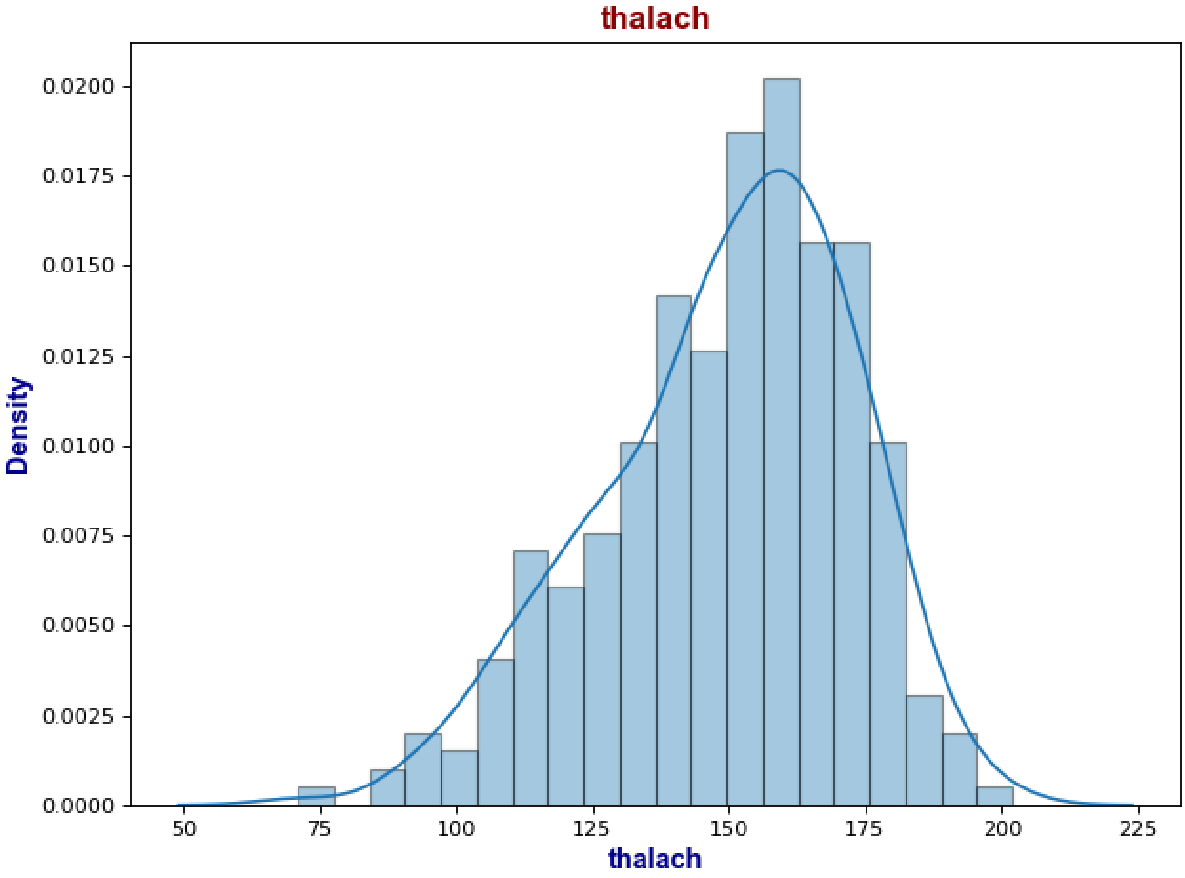


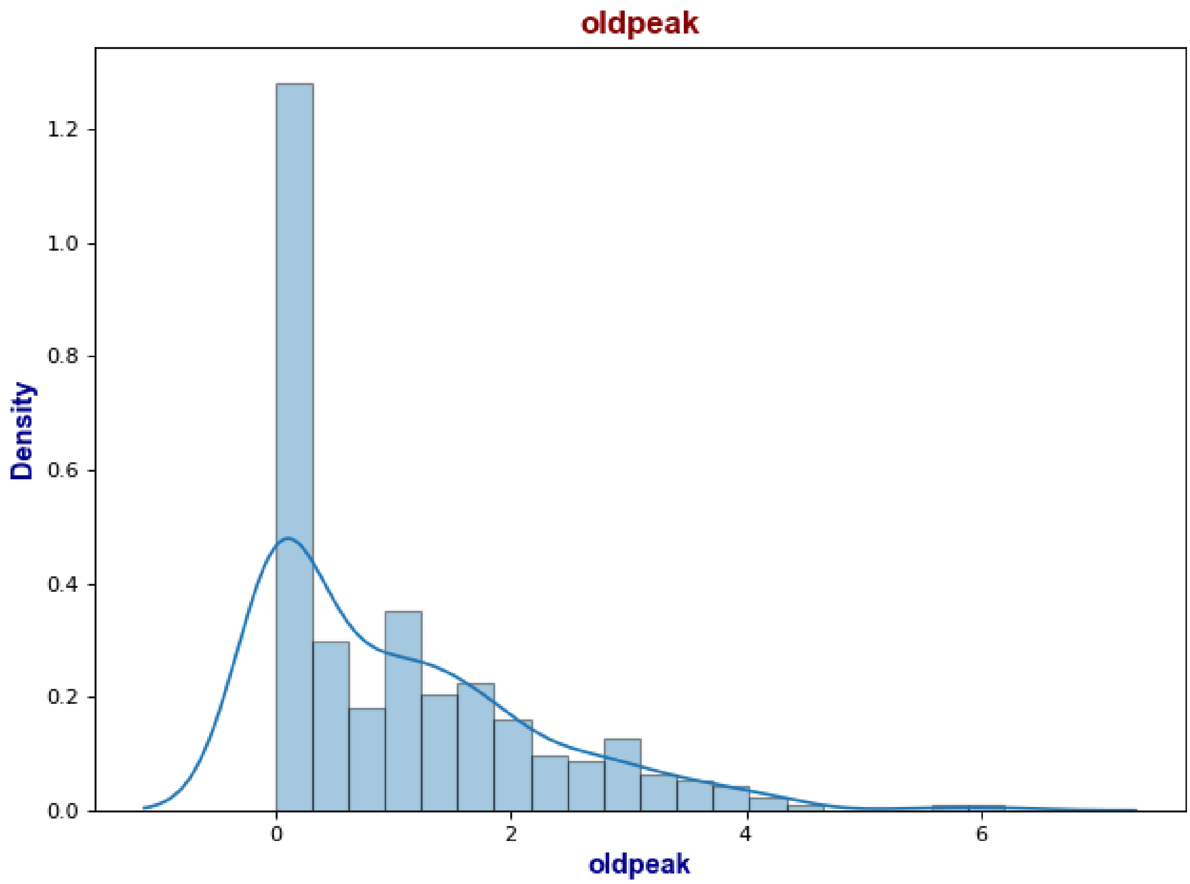
```
      Variable        VIF
0          age  39.567644
1          sex   3.507112
2           cp   2.409980
3     trestbps  58.776923
4         chol  26.281421
5          fbs   1.273256
6      restecg   2.051037
7     thalach   42.631809
8        exang   2.022825
9      oldpeak   3.071361
10       slope  10.015857
11          ca   1.860512
12        thal  17.141073
```

Images 2: Histograms of numerical variables

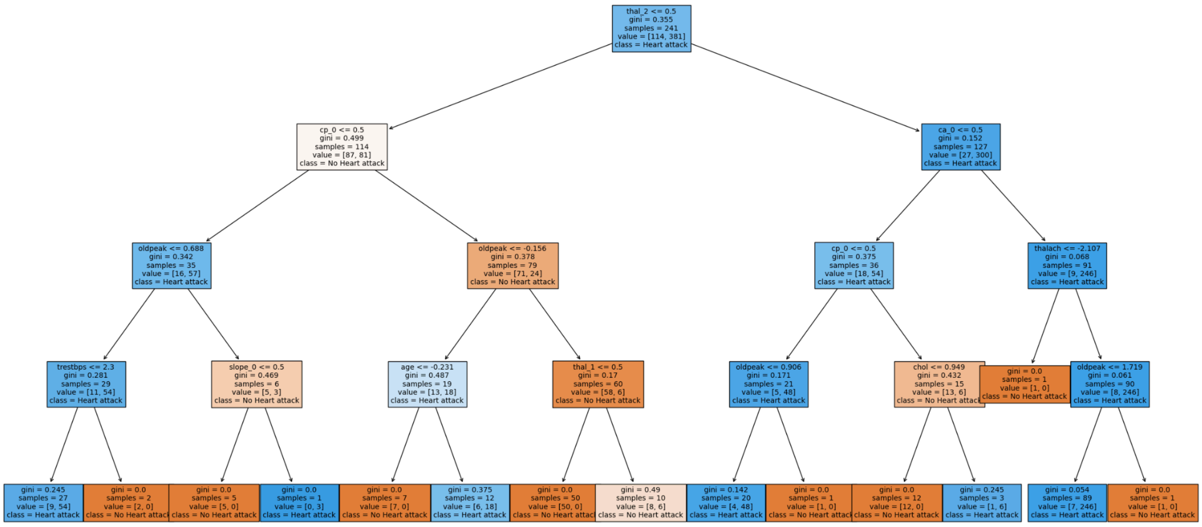



trestbps
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
80
100
120
140
160
180
200
220
trestbps

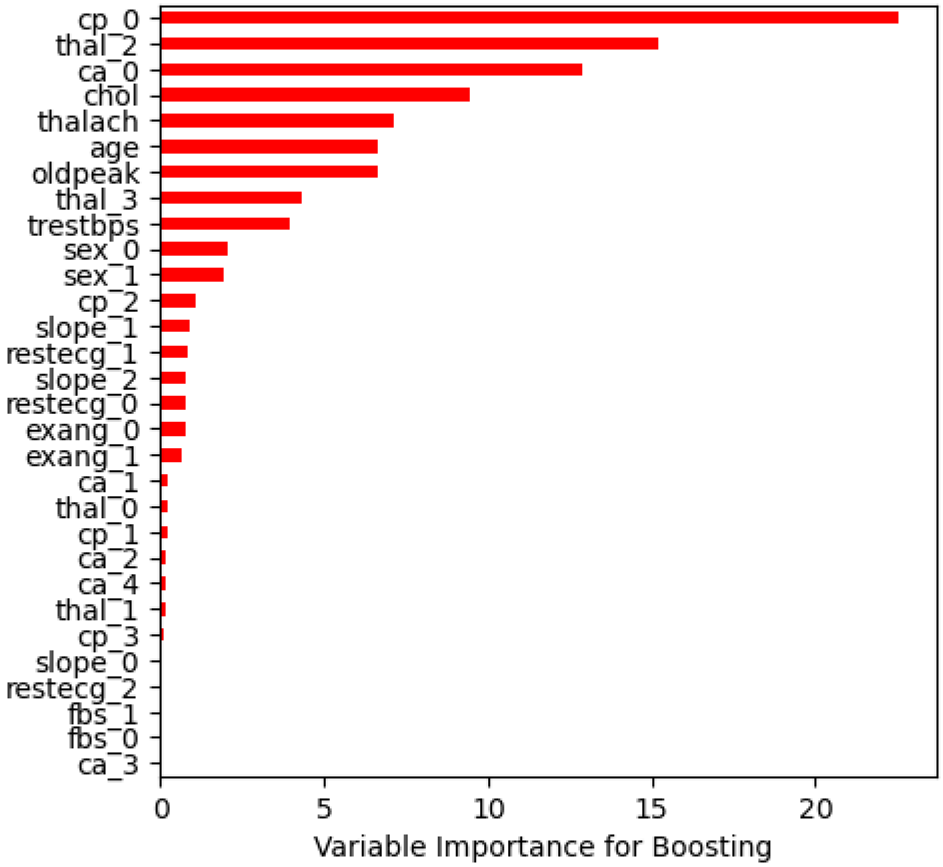

thalach
Density
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
50
75
100
125
150
175
200
225
thalach



Images 3 : Optimal Decision Tree



Images 4: Feature Important for Boosting


cp_0
thal_2
ca_0
chol
thalach
age
oldpeak
thal_3
trestbps
sex_0
sex_1
cp_2
slope_1
restecg_1
slope_2
restecg_0
exang_0
exang_1
ca_1
thal_0
cp_1
ca_2
ca_4
thal_1
cp_3
slope_0
restecg_2
fbs_1
fbs_0
ca_3
0
5
10
15
20
Variable Importance for Boosting